Use Anderson darling statistic to test the normality.

The data is:

88.9, 87, 90, 88.2, 87.4, 89.7, 87.8, 86, 89.6, 87.2

The mean (88.18) & SD (1.33) are given.

H_0: normal
H_1: not normal

$$A^2 = -n - \frac{1}{n}\sum\left\{(2i-1)\ln[\phi(z_i)] + [2(n-i)+1]\ln[1-\phi(z_i)]\right\}$$

When mean & SD are not given,

$$A^{*2} = A^2\left[1 + \frac{4}{n} - \frac{25}{n^2}\right]$$

Sol:

x_i	$z = \frac{x_i - 88.18}{1.33}$	cummulative $\phi(z)$ (-∞ to z)	i	$(2i-1)$ ①	$\ln(\phi(z))$ ②	①②
86	-1.64	0.0505	1	1	-2.9858	-2.9858
87	-0.89	0.1867	2	3	-1.6783	-5.0349
87.2	-0.74	0.2297	3	5	-1.4710	-7.355
87.4	-0.59	0.2776	4	7	-1.2816	-8.9712
87.8	-0.29	0.3859	5	9	-0.9522	-8.5698
88.2	0.02	0.5080	6	11	-0.6773	-7.4503
88.9	0.54	0.7054	7	13	-0.3490	-4.537
89.6	1.07	0.8577	8	15	-0.1535	-2.3025
89.7	1.14	0.8729	9	17	-0.1359	-2.3103
90	1.37	0.9147	10	19	-0.0892	-1.6948

$2(n-i)$	$2(n-i)+1$ ③	$1-\phi(z_i)$	$\ln[1-\phi(z_i)]$ ④	③④	①② + ③④
18	19	0.9495	-0.0518	-0.9842	-3.97
16	17	0.8133	-0.2067	-3.5139	-8.5488
14	15	0.7703	-0.2610	-3.9150	-11.27
12	13	0.7224	-0.3[illegible]252	-4.2276	-10.1988
10	11	0.6141	-0.4876	-5.3636	-13.9334
8	9	0.4920	-0.7093	-6.3837	-14.134
6	7	0.2946	-1.2221	-8.5547	-13.0917
4	5	0.1423	-1.9498	-9.7490	-14.286
2	3	0.1271	-2.0628	-6.1884	-8.4987
0	1	0.0853	-2.4616	-2.4616	-4.1564
					-102.0878

$$A^2 = -10 - \frac{1}{10}[-102.0878]$$

$$= -10 + 10.20878$$

$$= 0.2088$$

Multiple Regression Model:

$$\begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & Y \end{array}$$
$$\begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} & Y_1 \\ x_{12} & x_{22} & x_{32} & x_{42} & Y_2 \\ x_{13} & x_{23} & x_{33} & x_{43} & Y_3 \\ x_{14} & x_{24} & x_{34} & x_{44} & Y_4 \\ x_{15} & x_{25} & x_{35} & x_{45} & Y_5 \end{pmatrix}$$

Model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots \hat{\beta}_k X_k$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31} + \beta_4 x_{41} \\ \beta_0 + \beta_1 x_{12} \\ \beta_0 + \beta_1 x_{13} \\ \beta_0 + \beta_1 x_{14} \\ \beta_0 + \beta_1 x_{15} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & x_{41} \\ 1 & x_{12} & x_{22} & x_{32} & x_{42} \\ 1 & x_{13} & x_{23} & x_{33} & x_{43} \\ 1 & x_{14} & x_{24} & x_{34} & x_{44} \\ 1 & x_{15} & x_{25} & x_{35} & x_{45} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\boxed{Y = X\beta + \varepsilon}$$

$$Y_{5\times1} = X_{5\times5}\, \beta_{5\times1} + \varepsilon_{5\times1}$$

Parameter estimation.

$$\Sigma e_i^2 = e_1^2 + e_2^2 + e_3^2 + \cdots e_n^2 \qquad e = Y - \hat{Y}$$

$$e'e = \Sigma e^2 \qquad \Sigma e^2 = \Sigma (Y - \hat{Y})^2$$

$$Y = X\beta + \varepsilon$$

$$\hat{Y} = X\hat{\beta}$$

$$\varepsilon = Y - \hat{Y}$$

$$\varepsilon = Y - X\hat{\beta}$$

$$\varepsilon^2 = (Y - X\hat{\beta})^2$$

$$\varepsilon^2 = \varepsilon'\varepsilon$$

$$e'e = \varepsilon^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

$$= \left[Y' - (X\hat{\beta})'\right]\left[Y - X\hat{\beta}\right]$$

$$= \left[Y' - \hat{\beta}'X'\right]\left[Y - X\hat{\beta}\right]$$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - \hat{\beta}'X'Y - (\hat{\beta}'X'Y)' + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - \hat{\beta}'X'Y - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Note :
$\hat{\beta}'X'Y = (Y'X\hat{\beta})$
$e' = e$
$(Y'X\hat{\beta})' = Y'X\hat{\beta}$

Take Partial differentiation w.r to $\hat{\beta}$

$$\frac{\partial e'e}{\partial \hat{\beta}} = 0 - 2X'Y + 2X'X\hat{\beta}$$

Note:
$\hat{\beta}'X'X\hat{\beta} = \hat{\beta}'A\hat{\beta}$
where $X'X = A$
$\frac{\partial \hat{\beta}'A\hat{\beta}}{\partial \hat{\beta}} = 2A\hat{\beta}$

$$\frac{\partial e'e}{\partial \hat{\beta}} = 0$$

$$-2x'y + 2x'x\hat{\beta} = 0$$

$$2x'x\hat{\beta} = 2x'y$$

$$x'x\hat{\beta} = x'y$$

Note:

$xx^{-1} = I$

$x^{-1}x = I$

$x'x$ is matrix

$(x'x)^{-1}$ exists

Then Multiply both sides of the above equation by $(x'x)^{-1}$

$$(x'x)^{-1}(x'x)\hat{\beta} = (x'x)^{-1}x'y$$

$$I\hat{\beta} = (x'x)^{-1}x'y$$

$$\boxed{\hat{\beta} = (x'x)^{-1}x'y}$$

1. Properties of OLS estimators [ordinary least square]

$$\hat{\beta} = (x'x)^{-1}x'y$$

$$(x'x)\hat{\beta} = x'y$$

$$(x'x)\hat{\beta} = x'(x\hat{\beta} + e)$$

$$x'x\hat{\beta} = x'x\hat{\beta} + x'e$$

$$x'x\hat{\beta} - x'x\hat{\beta} = x'e$$

$$\boxed{0 = x'e}$$

$x_1'e = 0$

$x_1'e + x_2'e + \dots e_k'e = 0$

x_1	x_2	x_3	y
2	3	1	15
1	2	2	19
10	5	1	81

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$15 = \beta_0 + \beta_1 2 + \beta_2 3 + \beta_3 1$$

$$19 = \beta_0 + \beta_1 1 + \beta_2 2 + \beta_3 2$$

$$81 = \beta_0 + \beta_1 10 + \beta_2 5 + \beta_3 1$$

$$\underset{y}{\begin{bmatrix} 15 \\ 19 \\ 81 \end{bmatrix}} = \underset{x}{\begin{bmatrix} 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 10 & 5 & 1 \end{bmatrix}} \underset{\beta}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}} + \underset{\epsilon}{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}$$

$$y = x\beta + e$$

$$\hat{y} = x\hat{\beta}$$

$$e = y - \hat{y} = (y - x\hat{\beta})$$

$$\hat{\beta} = (x'x)^{-1}x'y$$

$$x' = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 10 \\ 3 & 2 & 5 \\ 1 & 2 & 1 \end{bmatrix}$$

$$x'x = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 10 \\ 3 & 2 & 5 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 10 & 5 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1+1+1 & 2+1+10 & 3+2+5 & 1+2+1 \\ 2+1+10 & 4+1+100 & 6+2+50 & 2+2+10 \\ 3+2+5 & 6+2+50 & 9+4+25 & 3+4+5 \\ 1+2+1 & 2+2+10 & 3+4+5 & 1+4+1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 13 & 10 & 4 \\ 13 & 105 & 58 & 14 \\ 10 & 58 & 38 & 12 \\ 4 & 14 & 12 & 6 \end{bmatrix}$$

$$|x'x| = \begin{vmatrix} 3 & 13 & 10 & 4 \\ 13 & 105 & 58 & 14 \\ 10 & 58 & 38 & 12 \\ 4 & 14 & 12 & 6 \end{vmatrix}$$

$$= 3\begin{vmatrix} 105 & 58 & 14 \\ 58 & 38 & 12 \\ 14 & 12 & 6 \end{vmatrix} - 13\begin{vmatrix} 13 & 58 & 14 \\ 10 & 38 & 12 \\ 4 & 12 & 6 \end{vmatrix} + 10\begin{vmatrix} 13 & 105 & 14 \\ 10 & 58 & 12 \\ 4 & 14 & 6 \end{vmatrix} - 4\begin{vmatrix} 13 & 105 & 58 \\ 10 & 58 & 38 \\ 4 & 14 & 12 \end{vmatrix}$$

$$= 3\left[105\begin{vmatrix} 38 & 12 \\ 12 & 6 \end{vmatrix} - 58\begin{vmatrix} 58 & 12 \\ 14 & 6 \end{vmatrix} + 14\begin{vmatrix} 58 & 38 \\ 14 & 12 \end{vmatrix}\right] -$$

$$13\left[13\begin{vmatrix} 38 & 12 \\ 12 & 6 \end{vmatrix} - 58\begin{vmatrix} 10 & 12 \\ 4 & 6 \end{vmatrix} + 14\begin{vmatrix} 10 & 38 \\ 4 & 12 \end{vmatrix}\right] +$$

$$10\left[13\begin{vmatrix} 58 & 12 \\ 14 & 6 \end{vmatrix} - 105\begin{vmatrix} 10 & 12 \\ 4 & 6 \end{vmatrix} + 14\begin{vmatrix} 10 & 58 \\ 4 & 14 \end{vmatrix}\right] -$$

$$4\left[13\begin{vmatrix} 58 & 38 \\ 14 & 12 \end{vmatrix} - 105\begin{vmatrix} 10 & 38 \\ 4 & 12 \end{vmatrix} + 58\begin{vmatrix} 10 & 58 \\ 4 & 14 \end{vmatrix}\right]$$

$$= 3\left[105[228-144] - 58[348-168] + 14[696-532]\right] -$$

$$13\left[13[228-144] - 58[60-48] + 14[120-152]\right] +$$

$$10\left[13(348-168) - 105[60-48] + 14[140-232]\right] -$$

$$4\left[13(696-532) - 105[120-152] + 58[140-232]\right]$$

$= 3\left[105(84) - 58(180) + 14(164)\right] - 13\left[13(84) - 58(12) + 14(-32)\right]$

$+ 10\left[13(180) - 105(12) + 14(-92)\right] - 4\left[13(164) - 105(-32) + 58(-92)\right]$

$= 3\left[8820 - 10440 + 2296\right] - 13\left[1092 - 696 - 448\right] +$

$10\left[2340 - 1260 - 1288\right] - 4\left[2132 + 3360 - 5336\right]$

$= 3\left[676\right] - 13\left[-52\right] + 10\left[-208\right] - 4\left[156\right]$

$= 2028 + 676 - 2080 - 624$

$= 0$

Properties of MLR [predicted]

1.) The observed values of x are uncorrelated with the residuals. $x'e = 0$

implies that for every column x_k of X ie, $x_k' e = 0$

In another words, each regressor has zero sample correlation with the residual. Note that this does not mean that x is uncorrelated with the disturbances, we will have to assume this

If our regression includes the constant then the following properties holds :

(2) Sum of residuals is zero

If there is a constant then the first column in x will be a column of 1's this means that for the first element in the $x'e$ vector are $(x_{11}e_1 + x_{12}e_2 + \dots + x_{1k}e_n) = 0$,

It must be the case that $\Sigma e_i = 0$

3.) The sample mean of the residuals is 0

$$\bar{e} = \frac{1}{n} \Sigma e_i$$

$$= \frac{1}{n}(0)$$

$$\boxed{\bar{e} = 0}$$

4.) The regression hyper plane passes through the means of the observed values ie.) $(\bar{x}, \bar{y})$

This follows from the fact $\bar{e} = 0$

$$e = Y - \hat{Y}$$

$$e = Y - X\hat{\beta}$$

$\div$ by n on both sides.

$$\frac{e}{n} = \frac{Y}{n} - \frac{X\hat{\beta}}{n}$$

Take summation on both sides

$$\frac{\Sigma e}{n} = \frac{\Sigma Y}{n} - \frac{\Sigma X\hat{\beta}}{n}$$

$$\bar{e} = \bar{Y} - \bar{X}\hat{\beta}$$

$$\boxed{0 = \bar{Y} - \bar{X}\hat{\beta}}$$

5.) The predicted values of Y are uncorrelated with the residuals.

$$\hat{Y} = X\hat{\beta}$$

$$\hat{Y}'e = (X\hat{\beta})'e$$

$$\hat{Y}'e = \hat{\beta}'X'e$$

$$\hat{Y}'e = \hat{\beta}'(0) \longrightarrow \text{From (1)},$$

$$\boxed{\hat{Y}'e = 0}$$

6.) The Mean of the predicted Y for the sample will equal to the mean of the observed Y

ie.) $\boxed{\hat{Y}' = \bar{Y}}$

We need to make some assumptions about the true model in order to make any inferences regarding β from $\hat{\beta}$. $\hat{\beta}$ comes from our sample but we want to learn about the true parameters.

<u>The Gauss Markov Assumptions</u>:

1.) $Y = X\beta + \varepsilon$

This assumption states that there is a linear relationship between Y and X.

2.) X is an $n \times K$ matrix of full rank. This assumption states that there is no perfect Multicollinearity. In other words, the columns of X are linearly independent this assumption is known as identification condition.

Note: if $|X| = 0$ (Multicollinearity) [dependent] singular matrix

$|X| \neq 0$ [independent] non singular matrix

We should transform the data

consider partial correlation coefficient,
(stepwise regression)

3.) $E\left[e/X\right] = 0$

$$E\begin{bmatrix} e_1/X \\ e_2/X \\ \vdots \\ e_n/X \end{bmatrix} = \begin{bmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_n] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

In other words, ~~[illegible]~~ no observation of the independent variables convey any information about expected value of disturbance.

4.) $E\left[\varepsilon\varepsilon'/x\right] = \sigma^2 I$

This assumption captures the familiar assumption of homoscedasticity and no auto correlation

$$E\left[\varepsilon\varepsilon'/x\right] = E\left\{\begin{pmatrix}\varepsilon_1/x \\ \varepsilon_2/x \\ \vdots \\ \varepsilon_n/x\end{pmatrix}\begin{pmatrix}\varepsilon_1/x & \varepsilon_2/x & \cdots & \varepsilon_n/x\end{pmatrix}\right\}$$

$$= \begin{bmatrix} E\left[\varepsilon_1/x\right]^2 & E\left[\varepsilon_1/x\,\varepsilon_2/x\right] & \cdots & E\left[\varepsilon_1/x\,\varepsilon_n/x\right] \\ E\left[\varepsilon_2/x\,\varepsilon_1/x\right] & E\left[\varepsilon_2/x\right]^2 & \cdots & \\ \vdots & E\left[\varepsilon_3/x\right]^2 & & \\ & & \cdots & E\left[\varepsilon_n/x\right]^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & & & & \\ & & & & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \\ \vdots & & & \ddots & \\ & & 0 & & 1 \end{bmatrix} = \sigma^2 I$$

$\Omega = \sigma^2 I$ where Ω is the variance covariance matrix of the disturbances.

$\Omega = E\left[\varepsilon\varepsilon'\right]$

5.) x may be fixed or random but must be generated by a mechanism that is not related to ε

6.) $\varepsilon/x \sim N(0, \sigma^2 I)$

Gauss Markov Theorem :-

The Gauss Markov theorem states that conditional on assumptions 1 to 5, there will be no other linear and unbiased estimator of the Beta coefficients that has a ~~[illegible]~~ smaller sampling variance. In other words, The ordinary least square [OLS] estimator is the best linear unbiased and efficient estimator. [BLUE].

$\hat{\beta}$ is an unbiased estimator of β.

(or)

prove: $E[\hat{\beta}] = \beta$

Sol:

We know that $\hat{\beta}$ matrix

$$\hat{\beta} = (x'x)^{-1} x'y$$

Model

$$y = x\beta + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_k x_k + \varepsilon$$

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$= (x'x)^{-1} x'(x\beta + \varepsilon)$$

$$= (x'x)^{-1} x'x\beta + (x'x)^{-1} x'\varepsilon$$

Taking Expectations on both sides,

Note: A is a matrix, $AA^{-1} = I$

$$E[\hat{\beta}] = E[I\beta + (x'x)^{-1} x'\varepsilon] \rightarrow (A)$$

$$= I\beta + (x'x)^{-1} * E[x'\varepsilon] \quad \text{where } x'\varepsilon = 0$$

$$= I\beta + (x'x)^{-1} * 0$$

$$\boxed{E[\hat{\beta}] = \beta}$$

case(i) x is fixed (non stochastic) so that we have

$$E[\hat{\beta}] = E[\beta + (x'x)^{-1} x'\varepsilon] \quad \text{from (A)}$$

$$= E[\beta] + E[(x'x)^{-1} x'\varepsilon]$$

$$= \beta + (x'x)^{-1} x' E[\varepsilon] \qquad E[\varepsilon] = 0, \; E[\varepsilon / x] = 0$$

$$= \beta + (x'x)^{-1} x' * 0$$

$$\boxed{E[\hat{\beta}] = \beta}$$

case(ii) x is a random or stochastic, we have

$$E[\hat{\beta}] = E[\beta + (x'x)^{-1} x'\varepsilon]$$

$$= E[\beta] + (x'x)^{-1} E[x'\varepsilon] \qquad E[x'\varepsilon] = 0$$

$$= \beta + 0$$

$$\boxed{E[\hat{\beta}] = \beta}$$

Variance of least square estimation:

W.K.T, $V(x) = E[x - E(x)]^2$

$$= E[(x - E(x))(x - E(x))']$$

$$\Rightarrow e^2 = ee'$$

$$var(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))']$$

where $\hat{\beta} = (x'x)^{-1} x'y$

$$= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

$$= E[((x'x)^{-1} x'y - \beta)((x'x)^{-1} x'y - \beta)']$$

Consider the term,

$$(X'X)^{-1}X'Y-\beta = (X'X)^{-1}X'(X\beta+\varepsilon)-\beta$$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon - \beta$$

$$= \cancel{I\beta} + (X'X)^{-1}X'\varepsilon - \cancel{\beta}$$

$$= (X'X)^{-1}X'\varepsilon$$

equation from (1),

$$Var[\hat{\beta}] = E\left[\left[(X'X)^{-1}X'\varepsilon\right]\left[(X'X)^{-1}X'\varepsilon\right]'\right]$$

$$= E\left[(X'X)^{-1}X'\varepsilon\right]\left[\varepsilon'X(X'X)^{-1}\right]$$

$$= (X'X)^{-1}X'X(X'X)^{-1}E[\varepsilon\varepsilon']$$

$$= (X'X)^{-1}(X'X)(X'X)^{-1}\sigma^2 I$$

$$= (X'X)^{-1} I \sigma^2 I$$

$$var(\hat{\beta}) = \sigma^2 (X'X)^{-1}_{p\times p}$$

Estimation of σ^2

$$SSE = \Sigma e^2$$

$$= e'e$$

$$= (Y-X\hat{\beta})'(Y-X\hat{\beta})$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - (\hat{\beta}'X'Y)' - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - \hat{\beta}'X'Y - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$Y'X\hat{\beta} = (\hat{\beta}'X'Y)' = \hat{\beta}'X'Y$ when it is symmetric

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$$

$$\boxed{SSE = Y'Y - \hat{\beta}'X'Y}$$

The mean sum of square is derived from SSE & its df

$$MSE = \frac{SSE}{n-p}$$, where $n \to$ no of observation, $p \to$ no. of parameter.

MSE is an unbiased estimate of σ^2.

Overall Performance of the model:

To check the overall performances, we use analysis of Variance

$$SST = SSR + SSE$$

where $\frac{SSR}{\sigma^2} = \chi^2_k$

$\frac{SSE}{\sigma^2} = \chi^2_{n-p}$

$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad (\beta_j = 0)$

$H_1 : \beta_1 \neq \beta_2 \neq \dots \neq \beta_k \neq 0 \quad (\beta_j \neq 0)$

Test Statistic:

$SST = \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$

$SSE = \Sigma Y_i^2 - \hat{\beta}' X' Y$

$SSR = \hat{\beta}' X' Y - \frac{(\Sigma Y_i)^2}{n}$

Construct of ANOVA table:

SV	df	SS	MSS	VR
Regression	K	$\hat{\beta}'X'Y - \frac{(\Sigma Y_i)^2}{n}$	$\frac{\hat{\beta}'X'Y - \frac{(\Sigma Y_i)^2}{n}}{k}$	$\frac{\hat{\beta}'X'Y - \frac{(\Sigma Y_i)^2}{n} / k}{\frac{\Sigma Y_i^2 - \hat{\beta}'X'Y}{n-p}} = F$
Error	$n-p$	$\Sigma Y_i^2 - \hat{\beta}'X'Y$	$\frac{\Sigma Y_i^2 - \hat{\beta}'X'Y}{n-p}$	
Total	$n-1$	$\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$		

k - no. of independent variables

p - no. of parameters

n - total no. of observation

Inference:

$F \Rightarrow$ The calculated value greater than table value $\Rightarrow$ Reject H_0

$F \Rightarrow$ The calculated value is less than table value $\Rightarrow$ Accept H_0

(or)

Reject H_0 if $F > F_{\alpha/2, k, n-p}$ df.

Inference: The estimated probability value is more than 0.05 then, Accept H_0 else Reject H_0.

Test for individual regression coefficient

To test the individual contribution of variables, To test the individual contribution of variables first frame the hypothesis as given below

$H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$, ..., $H_0 : \beta_k = 0$ $\rightarrow$ $H_0 : \beta_j = 0, \; j = 1, 2 \dots k$

$H_1 : \beta_1 \neq 0$, $H_1 : \beta_2 \neq 0$, ..., $H_1 : \beta_k \neq 0$ $\rightarrow$ $H_1 : \beta_j \neq 0, \; j = 1, 2, \dots k$

then take the level of significance (α)

Test Statistic

$$t_0 = \frac{\hat{\beta}_j - E[\hat{\beta}_j]}{SE[\hat{\beta}_j]}$$

$$= \frac{\hat{\beta}_j - \beta}{SE[\hat{\beta}_j]}$$

$SE[\hat{\beta}_j] = \sqrt{var(\hat{\beta}_j)}$ where $var[\hat{\beta}_j] = \sigma^2 (x'x)^{-1}_{p \times p}$

$= \sigma^2 C_{ij}$

thus

$$t_0 = \frac{\hat{\beta}_j - 0}{\sqrt{\sigma^2 C_{ij}}}$$

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{MSE\, C_{ij}}} \sim t_{n-p}$$

Inference:

Reject H_0 if $t_0 > t_{\alpha/2, n-p}$

when

$H_0 : \beta_j = c$

$H_1 : \beta_j \neq c$

$$t_0 = \frac{\hat{\beta}_j - c}{\sqrt{MSE\, C_{ij}}} \sim t_{n-p}.$$

Regression through origin / no intercept model

$Y_i = \beta_1 x_i + \varepsilon_i$

$\varepsilon_i = Y_i - \beta x_i$

$\varepsilon_i^2 = (Y_i - \beta_1 x_i)^2$

$S = \sum \varepsilon_i^2 = \sum (Y_i - \beta_1 x_i)^2$

Taking Partial differentiation w.r to β_1.

$$\frac{\partial S}{\partial \beta_1} = -2 \times \sum (Y_i - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow -2 \sum (Y_i - \beta_1 x_i) x_i = 0$$

$$\sum (Y_i - \beta_1 x_i) x_i = 0$$

$$\sum x_i Y_i - \beta_1 \sum x_i^2 = 0$$

$$\sum x_i Y_i = \beta_1 \sum x_i^2$$

$$\boxed{\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}}$$

$\hat{\beta}_1$ is an unbiased estimator of β_1

W.k.T $\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}$

Taking Expectation on both sides,

$$E[\hat{\beta}_1] = E\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right]$$

$$= \frac{\sum x_i E(Y_i)}{\sum x_i^2}$$

$$= \frac{\Sigma x_i \beta_1 x_i}{\Sigma x_i^2}$$

$$= \beta_1 \frac{\Sigma x_i^2}{\Sigma x_i^2}$$

$$\boxed{\hat{\beta_1} = \beta_1}$$

$$\mathrm{var}[\hat{\beta_1}] = \mathrm{var}\left[\frac{\Sigma x_i y_i}{\Sigma x_i^2}\right]$$

$$= \frac{\Sigma(x_i)^2 \, \mathrm{var}\, y_i}{(\Sigma x_i^2)^2}$$

$$= \frac{\Sigma(x_i)^2}{(\Sigma x_i^2)^2} \sigma^2$$

$$\boxed{\mathrm{var}[\hat{\beta_1}] = \frac{\sigma^2}{\Sigma x_i^2}}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon$$

$$Y = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & x_{32} & \cdots & x_{k2} \\ 1 & \vdots & & & & \\ \vdots & \vdots & & & & \\ 1 & x_{1n} & x_{2n} & x_{3n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2 + \cdots + \hat{\beta_k} x_k$$

$$e_i = Y_i - \hat{Y_i}$$

Mean error $= \frac{1}{n} \Sigma (Y_i - \hat{Y})$

Mean Absolute error $= \frac{1}{n} \Sigma |Y_i - \hat{Y}|$

$$\mathrm{MAPE} = \frac{1}{n} \Sigma \left| \frac{Y_i - \hat{Y}}{Y_i} \right| \times 100$$

where Y - response variable

$\hat{Y}$ - predicted value

n - total no. of observations.

Dummy Variable:

X - Explanatory variable.

(i) - numeric (Quantitative) ⇒ Regression Model

(ii) - Quality ⇒ Analysis of Variance

(iii) - numeric + Quality ⇒ Analysis of Covariance

Example 1: Different intercept & Same slope

$$D_2 = \begin{cases} 1 - \text{male} \\ 0 - \text{Female} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_2 + \varepsilon$$

$$E[\varepsilon] = 0 \; ; \; V[\varepsilon] = \sigma^2$$

when $D_2 = 1$

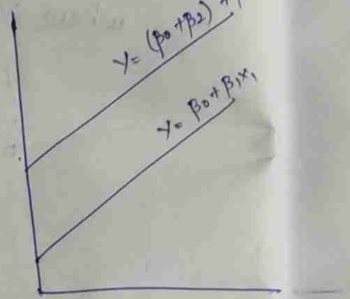
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (1)$$
$$= \beta_0 + \beta_1 X_1 + \beta_2$$

$$E[Y/D_2=1] = (\beta_0 + \beta_2) + \beta_1 X_1 \rightarrow (1)$$

when $D_2 = 0$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (0)$$

$$E[Y/D_2=0] = \beta_0 + \beta_1 X_1 \rightarrow (2)$$



(1) - (2),

$$E[Y/D_2=1] - E[Y/D_2=0] = (\beta_0 + \beta_2) + \beta_1 X_1 - \beta_0 - \beta_1 X_1$$
$$= \beta_2 \text{. is the coefficient of } D_2.$$

Example 2: Same intercept & different slope. (the line are parallel)

Suppose Y denotes the monthly salary of a person & D denotes whether the person is graduate or non graduate

Y - Salary

$$D - \text{Education} - \begin{cases} 1 - \text{Graduate} \\ 0 - \text{non Graduate} \end{cases}$$

The model is

$$Y = \beta_0 + \beta_1 \text{Education} + \varepsilon$$

$$Y = \beta_0 + \beta_1 D + \varepsilon$$

$$E[\varepsilon] = 0 \quad var[\varepsilon] = \sigma^2$$

The model equation when the person is graduated

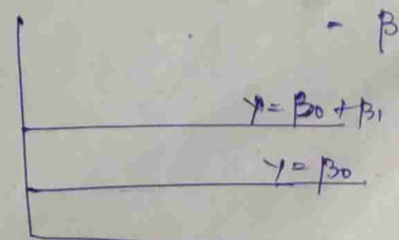
$$E[Y/D=1] = \beta_0 + \beta_1$$

The model equation when the person is nongraduate

$$E[Y/D=0] = \beta_0$$

The slope is defined as

$$E[Y/D=1] - E[Y/D=0] = [\beta_0 + \beta_1] - \beta_0$$
$$= \beta_1$$



Example 3

Let Y be considered as income; D_1 & D_2 are considered as education & gender respectively. Build all possible models with respect to present & absent of D_1, & D_2.

	D_1	D_2
$E[Y/D_1=0, D_2=0]$	0	0
$E[Y/D_1=1; D_2=0]$	1	0
$E[Y/D_1=0, D_2=1]$	0	1
$E[Y/D_1=1, D_2=1]$	1	1

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \varepsilon$$

$$E[\varepsilon] = 0 \quad Var[\varepsilon] = \sigma^2$$

$$E[Y/D_1=0, D_2=0] = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

$$E[Y/D_1=1, D_2=0] = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

$$E[Y/D_1=0, D_2=1] = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

$$E[Y/D_1=1, D_2=1] = \beta_0 + \beta_1(1) + \beta_2(1) = \beta_1 + \beta_1 + \beta_2$$

Dummy variable trap:

when $|A| = 0$ the proportion to the indicators
Drop β_0 & build model.

Different intercept & slope / Interaction term:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\text{Heart Attack} = \beta_0 + \beta_1(\text{BMI}) + \beta_2\overset{\text{yes/no}}{(\text{Sugar})} + \varepsilon$$

$$= \beta_0 + \beta_1(\text{BMI}) + \beta_2(\text{Sugar}) + \beta_3(\text{BMI})(\text{Sugar}) + \varepsilon.$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \beta_3 x_1 D_2 + \varepsilon$$

$$D_2 = \begin{cases} 1 - \text{Present (Sugar)} & - L_1 \\ 0 - \text{Absent (Sugar)} & - L_2 \end{cases}$$

when $D_2 = 0$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(x_1)(0)$$

$$\boxed{Y = \beta_0 + \beta_1 x_1} \rightarrow L_2$$

when $D_2 = 1$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(x_1)(1)$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1$$

$$\boxed{Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1} \rightarrow L_1$$

Multicollinearity:

$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k x_k + \epsilon$

Aim: Linear relationship b/w x & Y

$x_1, x_2, \ldots x_k$ are independent

$Y \leftarrow x_1, x_2 \ldots x_k$.

Model I:

$X_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \cdots + \beta_{k-1} x_k + \epsilon.$

$SST = SSR + SSE$

R^2

$VIF_1 = \frac{1}{1-R_i^2}$; $i=1$

Model II:

$x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots + \beta_{k-1} x_k + \epsilon$

$i=2$; $SST = SSR + SSE$

$VIF_2 = \frac{1}{1-R^2}$. $i=2$

Model k:

$x_k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_{k-1} x_{k-1} + \epsilon$

$SST = SSR + SSE$

$VIF_k = \frac{1}{1-R_i^2}$ $i=k$.

$VIF < 5$ [less multicollinearity]

$VIF \to 5 - 10$ (moderate)

$VIF > 10$ [severe]

To identify Multicollinearity Issue.

use (i) Ridge Regression

(ii) Lasso Regression

Assumption.

1.) Linearity

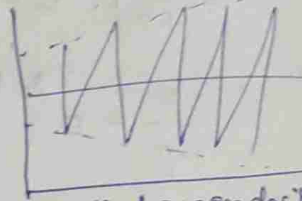
2.) $e \sim N(0, \sigma^2)$

3. detect outlier resolve

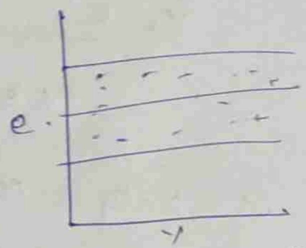
4. Multicollinearity $\to$ VIF

5. Residual Analysis

6. Homoscadacity.



Heteroscadacity



Homoscadacity.

Logistic Regression Analysis:

when data is imbalance, Accuracy is not best measure, we sensitivity & specificity.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

$$\hat{y} = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k}}$$

$$\text{Accuracy} = \frac{TP+TN}{N}$$

$$\text{Sensitivity} = \frac{\text{no. of 1's is correctly Predicted}}{\text{Total no. of 1's}}$$

$$\text{Specificity} = \frac{\text{no. of 0's correctly Predicted}}{\text{Total no. of 0's}}$$

when data is imbalance then try to balance the dataset,

3 methods: (i) sensitivity & specificity
(ii) ~~specificity~~ downsize.
(iii) upsize

Validation:

Data can be divided into Training data & Testing data [using simple random sampling without replacement].